Image Registration of Satellite Imagery with Deep Convolutional Neural Networks

Maria Vakalopoulou, Stergios Christodoulidis, Mihir Sahasrabudhe, Stavroula Mougiakakou, Nikos Paragios

IMAGE REGISTRATION OF SATELLITE IMAGERY WITH DEEP CONVOLUTIONAL NEURAL NETWORKS

Maria Vakalopoulou[1*] *Stergios Christodoulidis*[2*] *Mihir Sahasrabudhe*[1]
Stavroula Mougiakakou[2] *Nikos Paragios*[3]

[1]CVN, CentraleSupélec, Université Paris-Saclay and INRIA Saclay, France
[2]ARTORG Center, University of Bern, Bern, Switzerland, [3]TheraPanacea, Paris, France
{maria.vakalopoulou, mihir.sahasrabudhe}@centralesupelec.fr,
{stergios.christodoulidis, stavroula.mougiakakou}@artorg.unibe.ch, n.paragios@therapanacea.eu

ABSTRACT

Image registration in multimodal, multitemporal satellite imagery is one of the most important problems in remote sensing and essential for a number of other tasks such as change detection and image fusion. In this paper, inspired by the recent success of deep learning approaches we propose a novel convolutional neural network architecture that couples linear and deformable approaches for accurate alignment of remote sensing imagery. The proposed method is completely unsupervised, ensures smooth displacement fields and provides real time registration on a pair of images. We evaluate the performance of our method using a challenging multitemporal dataset of very high resolution satellite images and compare its performance with a state of the art elastic registration method based on graphical models. Both quantitative and qualitative results prove the high potentials of our method.

Index Terms— Deep Learning, Deformable and Linear Registration, Convolutional Neural Networks (CNN), Very High Resolution Satellite Images

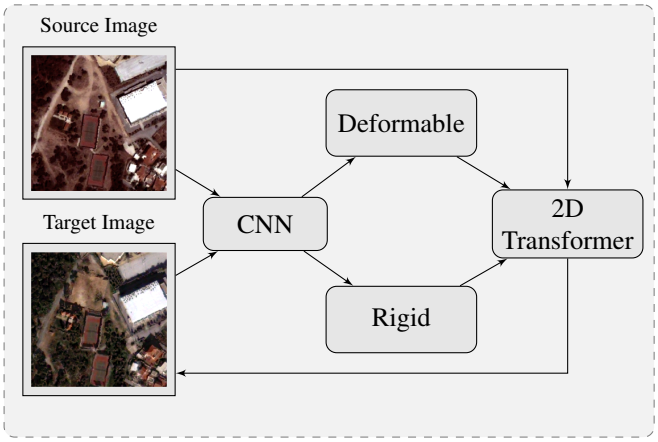

Fig. 1: A schematic diagram of the proposed framework. The input the pair of images is fed to a CNN architecture which predicts the rigid and deformable parameters for their registration. Then it uses a 2D spatial transformer layer to warp the source image and align it with the target.

1. INTRODUCTION

Given a pair of images depicting the same area, image registration is the process that aligns the source image S to the target image R. Numerous approaches have been proposed in order to address this problem and have been summarized in different surveys [1, 2]. Depending on the transformation model used, the methods can be categorized into two big groups: the rigid or linear and the deformable or elastic.

The rigid methods compute transformation models with mainly affine transformations e.g., rotation, scaling and translation. They are global in nature, thus, they cannot model local geometric differences between images. Numerous techniques [3, 4, 5] fall into this category and have been tested using different spectral and spatial resolution satellite imagery. On the other hand, deformable methods associate the observed pair of images through a non-linear dense transformation, or a spatially varying deformation model. These methods are commonly used in medical imaging and remote sensing datasets where the deformations between the images are not homogeneous [6, 7]. Non-linear transformations are more often observed in very high resolution satellite datasets from the same or different sensors, acquired from different angles as the displacements are not uniform.

Even though the problem of image registration is well-studied, there are a lot of challenges to be addressed, especially in remote sensing datasets. To start with, most methods in literature, especially the ones that calculate deformable transformations, are computationally quite expensive needing large time to compute the displacement field between two images. This is an important problem for remote sensing datasets in which images are larger. Moreover, accurate registration of multitemporal remote sensing imagery is very chal-

* Authors with equal contribution

lenging as the pair of images may contain areas which have changed or areas without changes but large differences in the recorded intensities. Depending on the acquisition time, the appearance of shadows can also make the accurate computation of displacement fields difficult. Even though there are methods that use similarity metrics that are not dependant on pixel intensities such as mutual information [8], and some that add supplementary tasks such as change detection [9] and/or semantic segmentation [10], the problem of registration still remains open.

In order to address the aforementioned challenges, in this paper we exploit a novel architecture which couples rigid and deformable registration to efficiently and accurately register remote sensing imagery (Figure 1). Our architecture is a modification of a recent work on accurate and efficient registration of 3D medical volumes [11]. In particular, the contributions of this paper are fourfold: *(i)* presenting a completely unsupervised convolutional neural network (CNN)-based registration framework, *(ii)* coupling rigid and deformable registration within a single optimization, *(iii)* presenting a framework which is independent of the CNN architecture, *(iv)* ensuring fast inference allowing real-time applications.

The rest of the paper is organised as follows. In Section 2, we present the proposed method. In Section 3, we describe the dataset and specify implementation details, while in Section 4 we discuss qualitative and quantitative results. We conclude in Section 5 and elaborate on possible future directions.

2. METHODOLOGY

The proposed framework can be divided into three different components—the transformation strategy, the CNN architecture, and the optimization procedure.

2.1. Linear and Deformable Transformer

The main component of the proposed CNN architecture is the 2D transformer layer, which warps the image S under a dense deformation G to create the warped image D,

$$D = \mathcal{W}(S, G), \tag{1}$$

where $\mathcal{W}(\cdot, G)$ indicates a sampling operation $\mathcal{W}$ under the deformation G.

The deformation is hence fed to the transformer layer as sampling coordinates for a backward bilinear interpolation sampling, adapting a strategy similar to [12]. The sampling process is then described by

$$D(\vec{p}) = \mathcal{W}(S, G)(\vec{p}) = \sum_{\vec{q}} S(\vec{q}) \prod_{d} \max\left(0, 1 - |[G(\vec{p})]_d - \vec{q}_d|\right), \tag{2}$$

where $\vec{p}$ and $\vec{q}$ denote pixel locations, $d \in \{x, y\}$ denotes an axis, and $[G(\vec{p})]_d$ denotes the d-component of $G(\vec{p})$.

The formulation we propose has two different components—one which calculates a linear/affine transformation A and another that calculates a dense transformation Φ. Depending on the application, these two terms can be used and trained together or separately. The transformation A corresponds to a 2×3 matrix, building a transformation grid G_A, which is the affine component of the deformation G.

For the deformable part G_N, we adopt an approach similar to [12]. Instead of regressing per-pixel displacements, we predict a matrix Φ of spatial gradients between consecutive pixels along each axis. As is discussed in [12], this approach helps generate smoother grids that render the deformable component easier to train. The actual grid G_N can then be obtained by applying an integration operation on Φ along x- and y-axes, which is approximated by the cumulative sum in the discrete case. We can then draw conclusions on the relative position of adjacent pixels in the warped image based on Φ. Concretely, two pixels $\vec{p}$ and $\vec{p} + 1$ will have moved closer, maintained distance, or moved apart in the warped image, if Φp is respectively less than 1, equal to 1, or greater than 1.

In case that the two parts are combined and trained together, the deformable grid G_N is applied first.

2.2. Network Architecture

Our formulation is independent of the network architecture and according to the application and dataset used, different ones can be incorporated. The architecture we used for our experiments is based on an encoder-decoder framework and it is very similar to the one presented in [13]. In particular, the encoder part adopts dilated convolutional kernels together with feature merging, while the decoder employs non-dilated convolutional layers. Specifically, a kernel size of 3×3 was set for the convolutional layers while LeakyReLU activation was employed for all convolutional layers. Each of the encoder-decoder parts contains 4 of these layer blocks with the feature maps starting from 16 and being doubled for each block, resulting in a 128 feature map. Before the decoder, all the feature maps were concatenated in order to create a more informative, multi-resolution feature space for the decoder. Finally, the decoder part has two different branches, one that calculates the affine parameters and one the deformable ones.

For the linear/affine parameters A, a linear layer was used together with a global average pooling to reduce the spatial dimensions, while for the spatial gradients Φ a sigmoid activation has been employed. Finally, the output of the sigmoid activation was scaled by a factor of 2 to allow consecutive pixels to have larger displacements than the initial.

2.3. Optimization

In our experiments, we used the mean squared error (MSE) between R and D to optimize our model using the Adam op-

timizer. The overall loss is defined as

$$\text{Loss} = \|R - \mathcal{W}(S, G)\|^2 + \alpha \|A - A_I\|_1 + \beta \|\Phi - \Phi_I\|_1 , \quad (3)$$

where A_I represents the identity affine transformation matrix, Φ_I the spatial gradients of the identity deformation, and α and β are regularization weights, The higher the values of α and β, the closer the deformation is to the identity. The regularization parameters are essential for the joint optimization, as they ensure that the predicted deformations will be smooth for both components. Moreover, the regularization parameters are very important in the regions of change, as they do not allow the deformations become very large.

3. DATASET AND IMPLEMENTATION DETAILS

For our experiments, we used a pair of multispectral very high resolution images from the Quickbird satellite. The pair has been acquired in 2006 and 2007, covering a $14\,\text{km}^2$ region in the East Prefecture of Attica in Greece. This particular dataset was challenging due to the very large size of the high resolution satellite images, their complexity due to different acquisition angles, shadows, important height differences, numerous terrain objects, etc. and the sparse multitemporal acquisitions. For evaluating the proposed architecture, patches of size 256×256 were created. In particular, 450 patches were selected randomly for training, 50 for validation and 50 for testing the proposed framework.

The initial learning rate was 10^{-3} and was divided by a factor of 10 if the performance on the validation set did not improve for 50 epochs while the training procedure stops when there is no improvement for 100 epochs. The regularization weights α and β were both set to 10^{-6}. For all the experiments we used a GeForce GTX 1080Ti GPU. We noted that the training converges after around 140 epochs. The overall training time was approximately 4 hours.

4. RESULTS AND DISCUSSION

To evaluate the performance of our method we perform different experiments using only the linear A or deformable Φ components, and also using their ensemble. Moreover, we compare the performance of our method with a state-of-the-art algorithm based on graphs as presented in [6] that has been proven to work very well on large remote sensing imagery. [6] used normalized cross correlation as the similarity metric.

Starting with the qualitative evaluation in Figure 2 we present three different pairs of images using checkerboard visualizations between the target R and warped image D before and after the registration using the different tested approaches. Even if the initial displacements were quite important all the methods recover the geometry and register the pair of images. However, the proposed method trained only with the A deformations fails to register accurately high buildings which have the largest deformations, due to the global nature of the transformation. Finally, the proposed method with only the deformable part, was slightly more difficult to be trained, proving that the additional linear component is a valuable part of the proposed framework.

Method	dx (pixel)	dy (pixel)	ds (pixel)	Time (sec)
Unregistered	7.3	6.3	9.6	–
Deformable [6]	1.3	2.3	2.6	$\sim$2
Proposed only A	2.5	2.8	3.7	$\sim$0.02
Proposed only Φ	1.2	2.0	2.3	$\sim$0.02
Proposed	**0.9**	**1.8**	**1.9**	$\sim$0.02

Table 1: Errors measured as average euclidean distances between estimated landmark locations. dx and dy denote distances along x-, y-, respectively, while ds denotes the average error along all axes.

Continuing with the quantitative evaluation, a number of landmarks, mainly on the buildings corners have been selected and their errors in each of the axes are reported in Table 1. It should be noted that for all the methods the same landmarks have been selected and around 10 image pairs were used to extract the landmarks. These landmarks contained mainly roofs of buildings as they were the ones presenting the higher registration errors. One can observe that the proposed method using only the affine transformation does not perform as well as the rest of the approaches as it fails to recover the geometry in places with local deformations. On the other hand the rest of the approaches report very low errors with the proposed method using both affine and deformable parts performing slightly better. Finally, it should be noted that the proposed method is very fast, with inference time for an image pair of size 256×256 less than half a second, giving a big advantage for very large datasets such as the remote sensing ones, and allowing even real-time applications.

5. CONCLUSION

In this paper we present a CNN-based method for the accurate registration of very high resolution images. The method is completely unsupervised, while it consists of two different parts, a linear and a deformable which can be trained together or separately. Our method reports accuracy similar to state-of-the-art methods with very small inference time. Our future steps include the extension of the formulation in a multi-task scheme, integrating loss functions that can handle better regions of change between the images. Extensions of the method to perform group-wise registration of more than two images simultaneously are also possible..

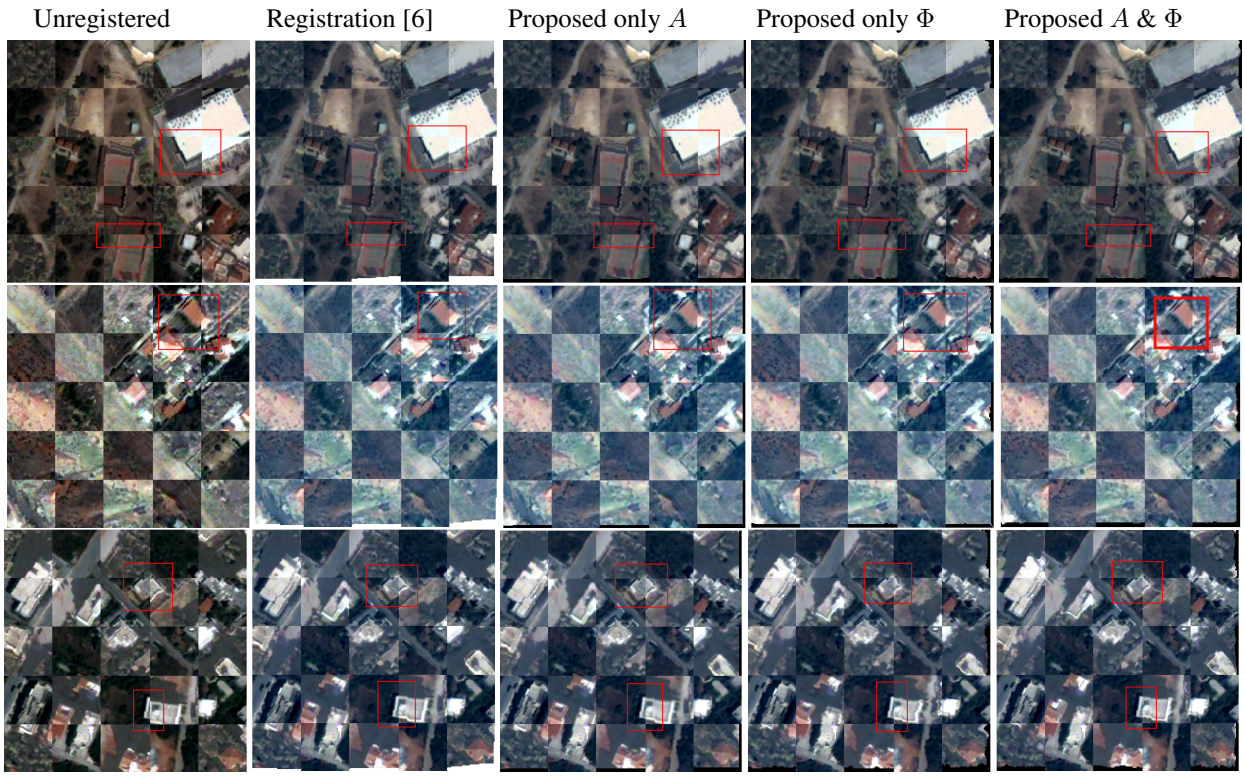

Fig. 2: Qualitative evaluation for three different pairs of images. With red rectangles we indicate regions of interest.

6. REFERENCES

[1] B. Zitova and J. Flusser, "Image registration methods: a survey," *Image and Vision Computing*, vol. 21, no. 11, 2003.

[2] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Transactions on Medical Imaging*, vol. 32, no. 7, 2013.

[3] J. Wu, C. Chang, H.-Y. Tsai, and M.-C. Liu, "C–registration between multisource remote-sensing images," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XXXIX-B3, 2012.

[4] M. Vakalopoulou and K. Karantzalos, "Automatic descriptor-based co-registration of frame hyperspectral data," *Remote Sensing*, vol. 6, no. 4, 2014.

[5] Z. Li and H. Leung, "Contour-based multisensor image registration with rigid transformation," in *2007 10th International Conference on Information Fusion*, 2007.

[6] K. Karantzalos, A. Sotiras, and N. Paragios, "Efficient and automated multimodal satellite data registration through mrfs and linear programming," in *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014.

[7] D. Marcos, R. Hamid, and D. Tuia, "Geospatial correspondences for multimodal registration," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] J. Zhang, M. Zareapoor, X. He, D. Shen, D. Feng, and J. Yang, "Mutual information based multi-modal remote sensing image registration using adaptive feature weight," *Remote Sensing Letters*, vol. 9, no. 7, pp. 646–655, 2018.

[9] M. Vakalopoulou, K. Karantzalos, N. Komodakis, and N. Paragios, "Graph-based registration, change detection, and classification in very high resolution multitemporal remote sensing data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 7, 2016.

[10] Lichao Mou, Xiao Zhu, Maria Vakalopoulou, Konstantinos Karantzalos, Nikos Paragios, Bertrand Le Saux, Gabriele Moser, and Devis Tuia, "Multitemporal very high resolution from space: Outcome of the 2016 ieee grss data fusion contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. PP, pp. 1–13, 06 2017.

[11] S. Christodoulidis, M. Sahasrabudhe, M. Vakalopoulou, G. Chassagnon, M-P Revel, S. Mougiakakou, and N. Paragios, "Linear and deformable image registration with 3d convolutional neural networks," in *Image Analysis for Moving Organ, Breast, and Thoracic Images*, 2018.

[12] Z. Shu, M. Sahasrabudhe, A. Guler Riza, D. Samaras, N. Paragios, and I. Kokkinos, "Deforming autoencoders: Unsupervised disentangling of shape and appearance," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[13] M. Anthimopoulos, S. Christodoulidis, L. Ebner, T. Geiser, A. Christe, and S. Mougiakakou, "Semantic segmentation of pathological lung tissue with dilated fully convolutional networks," *IEEE Journal of Biomedical and Health Informatics*, 2018.